Annals of the Institute of Statistical Mathematics (2024) 76:39–42
https://doi.org/10.1007/s10463-023-00885-3

INVITED ARTICLE: FOURTH AKAIKE MEMORIAL LECTURE

Discussion of "Identifiability of latent-variable and structural-equation models: from linear to nonlinear"

Takeru Matsuda[1,2]

Received: 23 June 2023 / Accepted: 4 July 2023 / Published online: 1 November 2023

First of all, I would like to celebrate Professor Hyvärinen for receiving the fourth Akaike memorial lecture award. He has developed many innovative methods for machine learning and signal processing, such as fastICA (Hyvärinen 1997), score matching (Hyvärinen 2005, 2007), noise contrastive estimation (Gutmann and Hyvärinen 2012), and Linear Non-Gaussian Acyclic Model (Shimizu et al. 2006, [LinGAM;][]). These ideas inspired many statisticians and led to various developments. For example, the score matching technique has been utilized in graphical models (Lin et al. 2016), Bayesian model comparison (Shao et al. 2019), proper scoring rules (Parry et al. 2012), and robust statistics (Yonekura and Sugasawa 2023). In the paper, Professor Hyvärinen provided a comprehensive survey of the identifiability theory for linear/nonlinear independent component analysis (ICA) models and structural equation models. It is written in an accessible form for statisticians and will be a nice guide for further statistical developments of the theory and application of nonlinear ICA and causal discovery.

The nonlinear ICA model is given by

$$x^{(k)} = f(s^{(k)}), \quad k = 1, \dots, K, \tag{1}$$

where $x^{(k)} \in \mathbb{R}^d$ is the data (observed), $s^{(k)} \in \mathbb{R}^d$ is the independent component (unobserved), and f is the mixing function (unknown). Note that $s_i^{(k)}$ and $s_j^{(k)}$ are independent for $i \neq j$. The goal is to estimate $s^{(1)}, \dots, s^{(K)}$ (and f) from $x^{(1)}, \dots, x^{(K)}$. If we only assume that $s^{(1)}, \dots, s^{(K)}$ are i.i.d., then this problem lacks identifiability (Hyvärinen and Pajunen 1999). Namely, even if we have infinite number of $x^{(k)}$, the

The Related Articles are https://doi.org/10.1007/s10463-023-00884-4; https://doi.org/10.1007/s10463-023-00886-2; https://doi.org/10.1007/s10463-023-00887-1.

✉ Takeru Matsuda
matsuda@mist.i.u-tokyo.ac.jp

[1] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

[2] RIKEN Center for Brain Science, 2-1 Hirosawa Wako City, Saitama 351-0198, Japan

function f cannot be uniquely determined. It can be understood by considering measure-preserving maps on the cube $[0,1]^d$.

For the nonlinear ICA model (1), Hyvärinen and Morioka (2016), Hyvärinen and Morioka (2017) showed that the identifiability is guaranteed by exploiting temporal structure. Later, Hyvärinen et al. (2019) extended this result to non-temporal structure as well. The key idea here is "learning via classification," where an unsupervised learning problem is translated to a self-supervised learning problem. Such an idea is also known as contrastive learning and used in several methods such as noise contrastive estimation (Gutmann and Hyvärinen 2012).

Technically, the above methods solve nonlinear ICA through training neural networks for classification tasks. Here, we briefly review classification with neural networks. Let x be data such as image and $z \in \{1, \dots, L\}$ be its label such as "dog" and "cat." Then, the classification probability by the softmax function is given by

$$p(z = l \mid x) = \frac{\exp(b_l + w_l^\top h(x))}{\sum_{j=1}^{L} \exp(b_j + w_j^\top h(x))}, \quad l = 1, \dots, L, \tag{2}$$

which can be viewed as the multinomial logistic regression from statistical viewpoint. The function $h(x)$ denotes the activation of units in the last hidden layer and the parameters $b = (b_1, \dots, b_L)$ and $w = (w_1, \dots, w_L)$ represent the bias and connection weights from the last hidden layer to the output layer, respectively. After training the network, the function h works well as a feature extractor for other data/tasks as well and it is often called the transfer learning.

In Time Contrastive Learning (Hyvärinen and Morioka 2016, [TCL;][]), the data $x^{(1)}, \dots, x^{(K)}$ are divided into consecutive segments (time windows) and a neural network is trained for segment classification. This is straightforwardly implementable with common deep learning packages. Hyvärinen and Morioka (2016) showed that the nonlinear independent components are obtained in the last hidden layer by TCL, under the assumption that the data are piecewise i.i.d. The intuition is: to attain better classification, better knowledge of the latent structure (nonlinear independent components) is necessary. This is similar in spirit to the generative adversarial networks (GAN). Hyvärinen and Morioka (2017) and Hyvärinen et al. (2019) extended the idea of TCL to more general cases.

Here, we revisit the above methods for nonlinear ICA from statistical perspective. From Bayes' rule, the softmax function (2) can be rewritten as

$$p(x \mid z = l) = g(x) \exp(w_l^\top h(x) - \psi(\theta)) \tag{3}$$

for some functions g and ψ, which means that the class distributions $p(x \mid z = 1), \dots, p(x \mid z = L)$ belong to the same exponential family (Efron 2022). Thus, training of neural networks for classification can be viewed as learning the nonlinearity h in the exponential family (3) in the last hidden layer. In this context, the result of Hyvärinen and Morioka (2016) can be interpreted as showing that the optimal nonlinearity h is (essentially) equal to the nonlinear independent components for the nonlinear ICA model (1) with piecewise i.i.d. structure. However, such an exponential family interpretation is not restricted to nonlinear ICA and it

may provide a theoretical basis for developing statistical methods that utilize (pre-trained) neural networks as feature extractors in general. For example, Matsuda and Hyvärinen (2019) developed a method for clustering unlabeled data that employs the feature from a pre-trained network and applies an extension of noise contrastive estimation to mixture models.

The idea of "learning via classification" has been utilized in several statistical methods as well. For example, bridge sampling (Bennett 1976; Meng and Wong 1996) is an algorithm for estimating the Bayes factor via classification of MCMC samples. Noise contrastive estimation (Gutmann and Hyvärinen 2012, [NCE;][]) can be viewed as an extension of bridge sampling that estimates both parameter θ and normalization constant $Z(\theta)$ in a non-normalized (energy-based) model

$$p(x \mid \theta) = \frac{1}{Z(\theta)}\widetilde{p}(x \mid \theta), \quad Z(\theta) = \int \widetilde{p}(x \mid \theta)\mathrm{d}x,$$

where the problem is reduced to classification of data and artificially generated noise. Note that Akaike information criterion has been extended to NCE as well as score matching (Matsuda et al. 2021). Other examples include the Cox regression (Cox 1972) and the semiparametric density ratio model (Qin 1998; Sugiyama et al. 2012).

References

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics, 22*, 245–268.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34*, 187–220.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Cambridge University Press.

Gutmann, M. U., Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research, 13*, 307–361.

Hyvärinen, A. (1997). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks, 10*, 626–634.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* 6, 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis, 51*, 2499–2512.

Hyvärinen, A., Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29*.

Hyvärinen, A., Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Workshop on Artificial Intelligence and Statistics*.

Hyvärinen, A., Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks, 12*, 429–439.

Hyvärinen, A., Sasaki, H., Turner, R. (2019). Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 22nd International Workshop on Artificial Intelligence and Statistics*.

Lin, L., Drton, M., Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics* 10, 806–854.

Matsuda, T., Hyvärinen, A. (2019). Estimation of non-normalized mixture models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Matsuda, T., Uehara, M., Hyvärinen, A. (2021). Information criteria for non-normalized models. *Journal of Machine Learning Research*, 307–361.

Meng, X. L., Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica, 6*, 831–860.
Parry, M., Dawid, P., Lauritzen, S. (2012). Proper local scoring rules. *Annals of Statistics, 40*, 561–592.
Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika, 85*, 619–630.
Shao, S., Jacob, P. E., Ding, J., Tarokh, V. (2019). Bayesian model comparison with the Hyvarinen score: Computation and consistency. *Journal of the American Statistical Association, 114*, 1826–1837.
Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research, 7*, 2003–2030.
Sugiyama, M., Suzuki, T., Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.
Yonekura, S., Sugasawa, S. (2023). Adaptation of the tuning parameter in general Bayesian inference with robust divergence. *Statistics and Computing, 33*, 39.